Michael Hughes
Modern Biology
F.W. Olin College of Engineering
December 2007

# Estimating Spontaneous Genetic Mutation Rates

## ABSTRACT

Spontaneous mutation causes both genetic novelty and genetic disorders. Scientists and medical professionals in fields from microbiology to epidemiology to taxonomy benefit from accurate predictions of mutation. Three distinct methods of estimating this rate are profiled below, each suited to a particular type of organism and research question. All techniques are literature grounded and used by biologists in actual applications.

The first technique measures differences in DNA within pseudogenes of related species and then leverages chronological knowledge of when the species diverged to estimate the mutation rate. The second observes that many genetic disorders (e.g. hemophilia) can arise from both heredity and mutation. Scientists can compare estimates for the likelihood that an affected individual will pass on the disease to children to census data of actual populations over time to approximate the number of mutation-induced cases of the disease occur over time. In the final method, an organism and its descendants can be sequenced and the mutation rate can be measured directly.

## 1. DEDUCTION USING SELECTION-NEUTRAL VARIANCE BETWEEN RELATED SPECIES

### *Key Idea*

Many genes in complicated eukaryotic genomes appear to never be transcribed. These so-called pseudogenes or “junk DNA” do not contribute to the organism’s survival or reproduction and are thus considered selection-neutral. Under no pressure from natural selection, pseudogene alleles vary purely according to random mutations. This can be true not only for a given population but for an entire species or even between closely-related species, depending on how far back the gene remains selection-neutral. By measuring variation among known pseudogenes shared by genetically-related species, the random genetic mutation rate of the younger species can estimated.

### *Calculation*

Using a model similar to that developed by Kimura [1], the sequenced genomes of individuals from two distinct species can be compared and a divergence score can be assigned indicating the amount of the two species' genomes have diverged since the younger species first evolved. To estimate a mutation rate which would yield this score, the number of generations since divergence and the size of the original divergent population are required. Once these parameters have been estimated using values from the literature, the actual mutation rate calculation is fairly straight-forward

$$\mu = \frac{k}{2t + 4N_e}$$

μ = mutation rate (mean number of mutations per nucleotide)
*k* = divergence score
*t* = time since the species have diverged (in generations)
*N*e = ancestral effective population size

*Research Example*

Nachman and Crowell [2] utilized this estimation technique in their 1999 calculation of human mutation rate based on pseudogene divergence from the chimpanzee genome. Comparison at 18 distinct loci, including 6 on the X chromosome, resulted in estimated rates within the range of $1.3x10^{-8}$ to $2.7x10^{-8}$ mutations per nucleotide site. The authors provide an average of ~$2.5x10^{-8}$ mutations per nucleotide site or 175 mutations per diploid genome per generation. These correspond to an assumed Homo-Pan divergence time between 4.5 and 6 million years ago and an initial divergent population between $10^4$ and $10^5$ individuals. Error associated with this calculation is much more likely attributed to difficulties in estimating the original population size or the divergence time rather than the calculation of variance, which is an exercise with a very low potential for error.

## 2. DEDUCTION FROM NON-HEREDITARY ONSET OF GENETIC DISEASE

*Key Idea*

Many heritable disorders within a wild population have been observed to occur with roughly constant frequency across generations. If a genetic disorder negatively impacts an individual's likelihood to reproduce, the consistent presence of the disorder in each successive generation must in some part be explained by the introduction of the disease into previously unaffected individuals via random mutation. The chance of this mutation happening can be estimated if reliable data on the effective reproduction rate of a diseased individual is known.

*Calculation*

First pioneered by Haldane in the 1920s and 1930s [3], this model requires the statistical calculation of a diseased individual's chance of producing offspring compared to an unaffected individual, a parameter known as effective fertility. This quantity, along with data regarding the number of affected individuals and the total number of individuals in the population, can be used to determine the number of diseased genes *W* which fail to propagate on to the next generation.

$$W = (1 - f)\, xN$$

The same number of genes must then be introduced in the next generation via random mutation. Knowing that the mutation may be possibly introduced at *l* distinct loci within a given individual, the mean rate *μ* at which a given allele will mutate is given by

$$W = l\mu N$$

Solving for μ gives

$$\mu = \frac{(1 - f)\, x}{l}$$

μ = mutation rate (mean number of mutations per allele)
*l* = number of unique loci which influence given disorder
*x* = proportion of population affected with given disorder
*f* = effective fertility of an affected individual

*Research Example*

One of the first detailed studies of mutation rate in humans was performed by Haldane [4]. Using available data regarding the population of male hemophiliacs in London, he obtained values for the frequency of hemophilia (estimated to range between $4x10^{-5}$ and $2x10^{6}$) and the effective fertility of hemophiliacs (between .1 and .25). Using these figures and setting the number of unique loci to 3 (2 for female, 1 for male since hemophilia is X-linked recessive), he computed that the mutation rate per locus ranges between $1x10^{-6}$ and $5x10^{-5}$ and provided a plausible average value of $2x10^{-5}$. This estimate is surprisingly accurate when considered against modern sequencing-based calculations, which place the mutation rate of many loci on the order of $10^{-5}$, including hemophilia. Haldane's efforts are even more impressive given the fact that he conducted his analysis well before widespread use of computers or even the "discovery" of DNA.

## 3. DIRECT MUTATION MEASUREMENT FROM DNA SEQUENCING OF DESCENDANTS

### *Key Idea*

Perhaps the most accurate and straightforward way to estimate a species' mutation rate is to directly observe and sequence several generations of organisms descended from a single ancestor whose genome is also known. Through comparison of sequences corresponding to various levels of ancestry, the exact generation in which a given mutation was first introduced can be determined, and an average error rate per generation can be computed from aggregated data of this nature. Comparison of these experimental "mutation accumulation" lines to wild type organisms also yields valuable results through similar analytical processes.

### *Calculation*

To determine the total number of mutations accrued in a given generation, powerful computational software is used to compare the experimental genome to some control genome (either a known ancestor genome or a normal wild-type genome) and record all variations. Once this number is determined, calculating the average mutation rate per nucleotide site per generation is very straightforward.

$$\mu = \frac{m}{LnT}$$

μ = mutation rate (per nucleotide site per generation)
m = number of observed mutations in a given generation
L = number of unique organisms in the given generation
n = number of nucleotide sites
T = generation number

### *Research Example*

The research of Denver et. al. [5] provides a thorough investigation into the mutation rate of an experimental population of C. elegans (a soil nematode worm). They created 198 distinct experimental "mutation accumulation" lines of worms, grown under optimal conditions to minimize the effects of natural selection. These lines were allowed to reproduce for hundreds of generations, with only one offspring per line kept at each generation. At three distinct stages throughout the experiment (corresponding to 280, 353, and 396 generations) about 60 to 70 of the lines were sequenced at random locations on all six possible chromosomes. The resulting sequences, each around 21 kb in size, were then directly compared to natural isolates of C. elegans. A total of 30 unique mutations were observed, yielding a mean average mutation rate of $2.1 \times 10^{-8}$ (+/- $7 \times 10^{-9}$) mutations per nucleotide site per generation. This result was around tenfold higher than previous estimates based on phylogenic analysis. Mutations were predominantly of the insertion-deletion variety (17/30), with deletions about 3 times more common than insertions.

# Works Cited

[5] Denver, Dee R., Krystalynne Morris, Michael Lynch, and W. Kelley Thomas. "High Mutation Rate and Predominance of Insertions in the Caenorhabditis Elegans Nuclear Genome." Nature 430 (2004): 679-682. 3 Dec. 2007 <http://hcgs.unh.edu/News/Denver2004.pdf>.

[6] Drake, John W., Brian Charlesworth, Deborah Charlesworth, and James F. Crow. "Rates of Spontaneous Mutation." Genetics 148 (1998): 1667-1686. 4 Dec. 2007 <http://www.genetics.org/cgi/content/full/148/4/1667>.

[2] Nachman, Michael W., and Susan L. Crowell. "Estimate of the Mutation Rate Per Nucleotide in Humans." Genetics 156 (2000): 297-304. 4 Dec. 2007 <http://www.genetics.org/cgi/content/full/156/1/297>.

[3] Haldane, J. B. S. A mathematical theory of natural and artificial selection. Part V. Selection and mutation. Proc. Camb. Philos. Soc. 23 (1927):838-844.

[4] Nachman, Michael W. "Haldane and the First Estimates of the Human Mutation Rate." Journal of Genetics 83 (2004): 231-233. 4 Dec. 2007 <http://www.iisc.ernet.in/academy/jgenet/Vol83No3/231.pdf>.

[1] Kimura, Motoo. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, United Kingdom. 1983.